中文語音文件自動摘要之摘要模型

陳怡婷 黃耀民 葉耀明 陳柏琳
台師大資工所 台師大資工所 台師大資教所 台師大資工所
g93470070@csie.ntnu.edu.tw angus@csie.ntnu.edu.tw ymyeh@ice.ntnu.edu.tw berlin@csie.ntnu.edu.tw

摘要

自動文件摘要依形成方式可分為摘錄式(Extractive)摘要與非摘錄式(Non-extractive or Abstract)摘要兩類。摘錄式摘要是依據設定之摘要比例從原文件中選出重要的句子、段落、或章節來組成摘要，非摘錄式摘要則是依文件內容中之主題或概念而重寫成的摘要。由於非摘錄式摘要方式仍具相當的困難度，因此目前自動文件摘要的相關研究技術大多以摘錄式摘要為主。常見的摘錄式摘要模型原則上可依據其特性分為逐字比對(Literal Term Matching)與概念比對(Concept Matching)二種方式，分別以向量空間模型(Vector Space Model, VSM)及潛藏語意分析(Latent Semantic Analysis, LSA)為代表。在論文中我們提出數種自動文件摘要模型。在逐字比對方式上提出隱藏式馬可夫模型；於概念比對的方式上，提出嵌入式潛藏語意分析模型(embedded LSA, eLSA)與主題混合模型(Topical Mixture Model, TMM)兩種摘要模型。我們在中文語音廣播新聞語料庫上實作了一系列的實驗，實驗結果均顯示使用所提出的三種摘要模型的摘要結果均較其它常見方法有顯著的提昇，同時嵌入式潛藏語意分析模型的結果亦優於原來的潛藏語意分析模型。

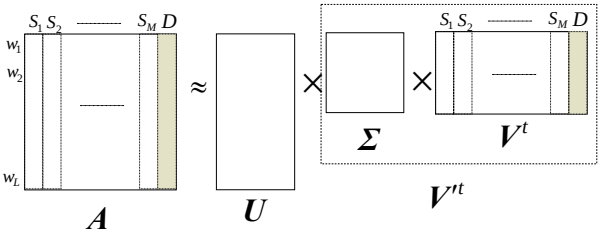
關鍵詞：自動文件摘要、隱藏式馬可夫模型、嵌入式潛藏語意分析模型、主題混合模型。

1. 前言

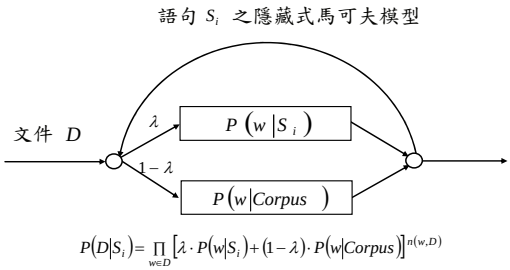
自動摘要的研究可追溯至 1950 年代晚期，並持續數十年至今。隨著網際網路的蓬勃發展，不僅使得自動文件摘要再度成為當前重要的研究範疇，同時亦延伸出許多新的研究方向，包含多文件摘要、多語言摘要及多媒體摘要等。自動文件摘要依形成方式可分為摘錄式(Extractive)與非摘錄式(Non-extractive or Abstract)摘要兩類。摘錄式摘要依據設定之摘要比例從原文件中選出重要的句子、段落、或章節來組成摘要，非摘錄式摘要則是依文件內容中之主題或概念而重寫成的摘要。由於非摘錄式摘要方式仍具相當的困難度，因此目前自動文件摘要的相關研究技術大多以摘錄式摘要為主[9]。

大多數的摘錄式自動摘要模型通常源於資訊檢索(Information Retrieval, IR)[11]。資訊檢索在於處理如何依使用者的查詢(Query)從大量的文件中找出相關（即符合使用者需求）文件(Documents)；而自動摘要則為找出與被摘要文件最相關或者最具有代表性的一至數個文句組合。例如常用於資訊檢索領域中的向量空間模型(Vector Space Model, VSM)[7]可以應用於自動摘要，其為將文件中每一文句及整篇文件均以一 L 維向量表示，向量的每一維度代表某個索引特徵(可以是詞、字或音節等單位)在文句或是文件中的統計值。每一文句分別依其與整篇文件之向量表示式的相關程度作排名，並依摘要比例摘錄出重要之文句。此外，若我們希望選取出重要的摘要文句並可以概括整篇文件的不同主題性，則可在每一次依序摘錄出某一重要文句後將文件中所有包含此文句中的索引特徵均予以刪除，再對其餘文句作相關程度排名，以減少摘錄出的文句間主題重覆性[14]，此方法稱為相關評估(Relevance Measure, RM)。另一種使用於概念比對的檢索模型－潛藏語意分析模型(Latent Semantic Analysis, LSA)[8]亦可被應用在文件摘要上，潛藏語意分析模型先將文件以“索引－文句”矩陣表示，然後透過奇異值分解(Singular Value Decomposition, SVD)將文件投射到一低維度的潛藏語意空間，並假設每一奇異值及其對應的奇異向量(Singular Vector)代表一概念(值越大越重要)，且文件中每一文句可由右奇異矩陣轉置的行向量表示。接著，依所對應奇異值大至小，從右奇異向量(右奇異矩陣轉置的列向量)取出最大索引值的文句作為文件的摘要[14]。其他摘要方法像是將文件中的每一文句以一連串索引特徵(例如詞或音節等)表示，並以文句中各索引特徵的統計值(如詞頻、反文件頻等統計資訊)及語言評估值(如對類專有名詞或是不同詞性的詞給予不同的分數)加權後累加值，作為文句的重要性分數，並以此分數做為文句選取的依據[10]。除上述所描述的選取重要文句的方法外，亦可進一步對於重要文句進行縮減，像是刪除文句中無意義或不重要的索引特徵。

上述所提之摘要方法均可被應用於文字內容文件及語音文件上。但是，對於語音文件的摘要處理上存在著其他困難點，像是語音辨識所產生的辨識錯誤及無法正確分辨句子或段落邊界的問題。為了避免選取重要資料時含有多餘或不正確的部分，於句子重要性計算時加入了許多其他的考慮因素，如語音辨識信心度分數、語言模型分數及文句文法等可利用資訊[13]。同時，一些額外的語音特徵如聲調、節奏、停頓時間等亦可視為重要的摘要特徵，不過此方面的特徵尚未被研究出一套可靠又有效率的整合方式。就語音文件摘要的呈現方式而言，可分為文字或語音訊號兩種呈現形式。前者優點在於方便瀏覽及進一步應用，但是無法避免語音辨識錯誤的影響，並且無法保留語音訊號所特有的聲韻資訊，像是說話者的語氣、音仰頓挫等；而後



圖一、嵌入式潛藏語意分析模型。



圖二、隱藏式馬可夫模型。

者在呈現時雖不會受語音辨識錯誤的影響，但是在組合所摘錄出的語句時，會有不自然或不流暢的問題需面對[13]。

本論文我們提出數種自動文件摘要模型。在逐字比對的方式上提出以隱藏式馬可夫模型應用於自動摘要；在概念比對的方式上提出嵌入式潛藏語意分析模型與主題混合模型兩種摘要模型。並於中文語音廣播新聞語料庫上實作了一系列的實驗，實驗結果均顯示所提出的摘要模型在語音文件自動摘要上均有不錯的表現。

本論文接下來的安排如下：第二節將介紹我們所提出的摘要模型與訓練方式；第三節將呈現相關的實驗設定、實驗結果及分析，第四節為結論及未來展望。

2. 文件自動摘要模型

本論文所提出之嵌入式潛藏語意分析模型、隱藏式馬可夫模型、主題混合模型等自動摘要模型，將在下面各小節中分別作介紹。

2.1 嵌入式潛藏語意分析模型

我們提出的嵌入式潛藏語意分析模型是將被摘要的文件中每一文句與文件本身共同投影到一個潛藏語意空間，最後藉由向量空間的餘弦評估方式，估測每一文句與整篇文件的相關性。在實作時可透過下列步驟完成：

(I) 將被摘要文件 D 斷句成數個文句，$D=S_1,S_2,..S_i..,S_M$。

(II) 利用每篇文件中所有文句建立起"索引-文句"的矩陣 $\boldsymbol{A}$，其中索引特徵可以是詞、字、音節或其他組合。並將整篇文件嵌入到矩陣的最後一行，如圖一所示。

(III) 對矩陣 $\boldsymbol{A}$ 進行奇異值分解，將其投射到潛藏語意空間，得到左奇異向量矩陣 $\boldsymbol{U}$、奇異值矩陣 $\boldsymbol{\Sigma}$（對角矩陣）與右奇異向量矩陣轉置 $\boldsymbol{V}^t$。

(IV) 在右奇異向量矩陣轉置 $\boldsymbol{V}^t$ 中，最後一行向量即為整篇文件於潛藏語意空間的向量表示式，其餘行向量即為原始文件中各文句在此空間的向量表示式。將 $\boldsymbol{\Sigma}$ 與 $\boldsymbol{V}^t$ 相乘可得到所有文句與整篇文件對不同維度加權過後向量表示式($\boldsymbol{V}'^t = \Sigma \times \boldsymbol{V}^t$)。

(V) 將 $\boldsymbol{V}'^t$ 的最後一行向量(即整篇文件的向量表示式)與 $\boldsymbol{V}'^t$ 中的其他行向量(各文句的向量表示式)進行餘弦相關度估測，得到每一文句之排名。

(VI) 依摘要比例，將句排名所對應的文句，摘錄形成摘要。

如圖一所示，$\boldsymbol{A}$ 與 $\boldsymbol{V}^t$ 的最後一行向量分別代表整篇文件在原來向量空間與潛藏語意空間的表示式。

2.2 隱藏式馬可夫模型

我們過去將隱藏式馬可夫模型用於中文語音資訊檢索(Spoken Document Retrieval, SDR)，獲得不錯的成果[2]；在本研究，我們將其延伸應用至語音文件自動摘要。隱藏式馬可夫模型將每篇文件 D 中每一文句 S_i 視為一個機率生成模型，如圖二所示。每一文句 S_i 對於每一個索引(可以是詞或音節，或是數個詞或音節的組合等)都可產生一個機率值。被摘要的文件與此文句的相關程度，是藉由文件中所有的索引在文句 S_i 發生可能性(Likelihood)來決定。文句 S_i 可以看成有兩個狀態(States)，分別對應到兩個機率分佈，其中 $P(w|S_i)$ 為索引 w 在文句 S_i 發生的機率；$P(w|Corpus)$ 為索引 w 在整個文件集(或語料)發生的機率，是用來平滑化(Smooth)索引 w 在文句 S_i 發生的機率，並同時表示索引 w 在語言裡的統計資訊。我們可以進一步假設，當給定一個文句模型時，文件 D 的每一個索引彼此是獨立的。因此，當文件中的索引在此文句的機率分佈值連乘積越高，則文件與此文句的相關度就越高，其公式如下：

$$P(D|S_i)=\prod_{w\in D}\left[\lambda\cdot P(w|S_i)+(1-\lambda)\cdot P(w|Corpus)\right]^{n(w,D)} \quad (1)$$

其中 $n(w,D)$ 是索引 w 在 D 出現的次數；λ 是比重參數(Weighting Parameter)。對於參數 λ 與每一文句 S_i 產生各索引 w 的機率值 $P(w|S_i)$，可以透過期望值最大化(Expectation-Maximization, EM)訓練[6]自動調整，模型訓練公式可表示如下：

$$\hat{\lambda}=\frac{\sum_{w\in D}E(w,S_i)}{\sum_{w'\in D}n(w',D)} \quad (2)$$

$$\hat{P}(w|S_i)=\frac{E(w,S_i)}{\sum_{w'\in D}E(w',S_i)} \quad (3)$$

$$E(w,S_i)=n(w,D)\frac{\lambda\cdot P(w|S_i)}{\lambda\cdot P(w|S_i)+(1-\lambda)\cdot P(w|Corpus)} \quad (4)$$

以隱藏式馬可夫模型來從事文件摘要的步驟可歸

納如下：

(I) 將被摘要文件 D 斷句成數個文句，$D=S_1,S_2,\ldots S_i\ldots,S_M$。

(II) 對文件 D 中每一文句 S_i，計算其單連語言模型 $P(w|S_i)$ 與比重參數 λ。

(III) 對文件 D 中各文句 S_i 估測機率值 $P(D|S_i)$，並依此作為文句排名。

(IV) 依摘要比例將排名結果所對應的文句輸出形成摘要。

此外，由於文件 D 中包含有文句 S_i 的資訊，所以可於估測文件與此文句的相關度時先去除文件 D 中所含文句 S_i 的相關資訊，我們稱之為文句移除(Sentence Removal, SR)。故式(1)可以進一步表示成：

$$P(D|S_i)=\prod_{w\in(D-S_i)}\left[\lambda\cdot P(w|S_i)+(1-\lambda)\cdot P(w|Corpus)\right]^{n(w,D)} \quad (5)$$

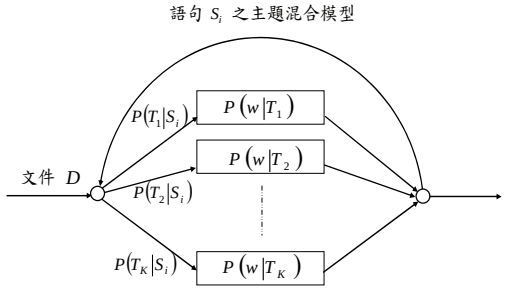
2.3 主題混合模型

主題混合模型最早由[1]所提出並且使用於語音資訊檢索，在此我們則嘗試將它延伸用於語音文件自動摘要。給定一篇將被摘要的文件 D，我們同樣地可以將 D 中每一文句 S_i 視為一個機率生成模型。不同的是，在此我們將每一文句 S_i 視為一個包含了 K 個潛藏主題的混合模型(Mixture Model)，其中每一潛藏主題 T_k 分別由一個單連語言模型 $P(w|T_k)$ 所表示，而且每一文句 S_i 對於每一潛藏主題 T_k 有不同的權重 $P(T_k|S_i)$，如圖三所示。我們可以進一步假設，當給定一個文句模型時，文件 D 的每一個索引彼此是獨立的。則文件 D 與每一文句 S_i 的相關程度可進一步地表示為：

$$P(D|S_i)=\prod_{w\in D}\left[\sum_{k=1}^{K}P(w|T_k)P(T_k|S_i)\right]^{n(w,D)} \quad (6)$$

由式(6)可看出，不像在隱藏式馬可夫模型採用逐字比對方式(Literal Term Matching)，文件 D 與每一文句 S_i 的相關程度取決於 $P(w|S_i)$，也就是文件 D 中的索引出現在文句 S_i 的機率；在主題混合模型，文件 D 與每一文句 S_i 的相關程度取決於 D 中的索引出現在某一個主題的機率 $P(w|T_k)$ 以及文句 S_i 產生此主題的機率 $P(T_k|S_i)$，當兩者的乘積越大代表文件 D 與文句 S_i 愈相關，即使文件 D 中的大部分索引並未出現於文句 S_i 中。因此，使用主題混合模型可以達到概念比對(Concept Matching)的目的。

另一方面，由於我們可以從網絡上得到與被摘要語音文件(例如廣播新聞)同一時期(Contemporary)或同一領域(In-domain)的文字新聞文件集 $\{D_c\}$，這些新聞文件通常都會有一句人工產生的標題，它其實就是一個非常簡潔的文件摘要資訊，於是我們可以將上述文件集中每篇文件 D_c 與其標題 H_c 對應關係 (D_c,H_c) 作為我們所提出之主題混合模型的訓練使用。同樣地，我們可以將每篇文字新聞文件的標題視為一個主題混合模型用來產生文件本身：



$$P(D|S_i)=\prod_{w\in D}\left[\sum_{k=1}^{K}P(w|T_k)P(T_k|S_i)\right]^{n(w,D)}$$

圖三、主題混合模型。

$$P(D_c|H_c)=\prod_{w\in D_c}\left[\sum_{k=1}^{K}P(w|T_k)P(T_k|H_c)\right]^{n(w,D_c)} \quad (6)$$

其中 $n(w,D_c)$ 是索引 w 在文件 D_c 出現的次數，我們可以假設文字新聞文件的標題模型間共用相同的主題機率分佈 $P(w|T_k)$，而各自對於每一潛藏主題 T_k 有不同的權重 $P(T_k|H_c)$。於是，我們可以透過期望值最大化(EM)訓練來估測標題模型的模型參數模型訓練公式可用下列式子表示：

$$\hat{P}(w|T_k)=\frac{\sum_{D_c\in\{D_c\}}n(w,D_c)P(T_k|w,H_c)}{\sum_{D_c\in\{D_c\}}\sum_{w'\in D_c}n(w',D_c)P(T_k|w',H_c)} \quad (7)$$

$$\hat{P}(T_k|H_c)=\frac{\sum_{w\in D_c}n(w,D_c)P(T_k|w,H_c)}{\sum_{w\in D_c}n(w,D_c)} \quad (8)$$

$$P(T_k|w,H_c)=\frac{P(w|T_k)P(T_k|H_c)}{\sum_{l=1}^{K}P(w|T_l)P(T_l|H_c)} \quad (9)$$

其中，$P(T_k|w,H_c)$ 為在索引 w 與標題 H_c 出現的條件下潛藏主題 T_k 發生的機率。在從事語音文件摘要時，對於被摘要文件 D 的每一文句 S_i 的主題混合模型，我們可以將式(7)所得 $\hat{P}(w|T_k)$ 用於式(6)的 $P(w|T_k)$，而式(6)中的 $P(T_k|S_i)$ 同樣地也可以透過期望值最大化(EM)訓練，使用式(8)及式(9)的修改來估測:

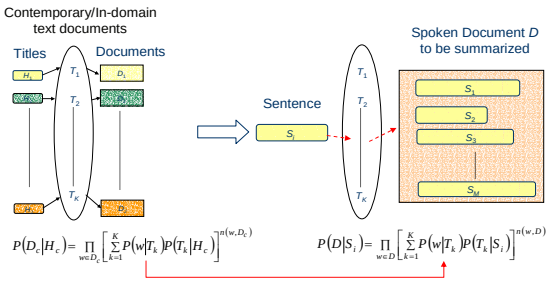
$$\hat{P}(T_k|S_i)=\frac{\sum_{w\in S_i}n(w,S_i)P(T_k|w,S_i)}{\sum_{w\in S_i}n(w,S_i)} \quad (10)$$

$$P(T_k|w,S_i)=\frac{P(w|T_k)P(T_k|S_i)}{\sum_{l=1}^{K}P(w|T_l)P(T_l|S_i)} \quad (11)$$

在以主題混合模型進行文件摘要時，我們可以同時把文件中索引在每一文句 S_i 中的機率分佈考慮進來，而將式(6)進一步表示成：

$$P(D|S_i)=\prod_{w\in D}\left[\lambda\cdot P(w|S_i)+(1-\lambda)\cdot\left(\sum_{k=1}^{K}P(w|T_k)P(T_k|S_i)\right)\right]^{n(w,D)} \quad (12)$$

其中 λ 是比重參數。同樣地對於參數 λ 與每一文句 S_i 產生各索引 w 的機率值 $P(w|S_i)$，可使用式(2)、式(3)自動調整，並修改式(4)為式(13)後：



圖四、語音文件摘要使用同一時期(同一領域)文字文件標題主題混合模型與被摘要文件文句主題混合模型。

$$E(w,S_i)=n(w,D)\frac{\lambda\cdot P(w|S_i)}{\lambda\cdot P(w|S_i)+(1-\lambda)\cdot\left(\sum_{k=1}^{K}P(w|T_k)P(T_k|S_i)\right)} \quad (13)$$

使用主題混合模型來從事文件摘要的步驟可歸納如下：

(I) 以同一時期或同一領域的文字新聞文件集(包括文件與標題對應關係)估測機率值 $P(w|T_k)$。

(II) 將被摘要文件 D 斷句成數個文句，$D=S_1,S_2,..S_i..,S_M$。

(III) 對文件 D 中每一文句 S_i 估測機率值 $P(T_k|S_i)$。

(IV) 對文件 D 中各文句 S_i 估測機率值 $P(D|S_i)$，並依此為作文句排名。

(V) 依摘要比例將排名結果所對應的文句輸出形成摘要。

最後，我們可以用圖四來表示語音文件摘要使用同一時期或同一領域文字文件之標題主題混合模型與被摘要文件之文句混合模型間關係。

3. 實驗評估

3.1 實驗語料

實驗語料蒐集自 News 98 新聞網 2001 年 8 月 1 日至 8 月 24 日中午 12:00 到 13:00 的 FM 廣播新聞，共 200 則廣播新聞，分為自動轉寫(Automatic Transcription)與人工轉寫(Manual Transcription)兩部分，相關統計資料如表一所示[15]。自動轉寫部分包含兩種語音辨識結果，相關資訊如表二所示[4]。隱藏式馬可夫模型與主題混合模型所需要的同一時期或同一領域文字文件訓練語料庫，用以訓練得到語言模型 $P(w|Corpus)$ 及"文件－標題"(D_c,H_c) 的訓練範例集，是採用中央通訊社(Central News Agency, CNA)在西元 2001 年 08 月所發佈且型態屬於故事(Type="story")的文字新聞做為文字文件訓練語料庫[16]，每一篇新聞皆含有文件與標題兩部份，其內容包括國內外及大陸文教、交通、社會、

表一、News 98 廣播新聞之語音文件相關統計資訊。

新聞時間	2001 年 8 月 1 日~ 2001 年 8 月 24 日
新聞數	200 則
總長度	1.61 小時
平均每則新聞長度	28.96 秒
總大小(人工轉寫)	約 31 仟字
平均每則新聞大小(人工轉寫)	約 157 字

表二、語音文件自動轉寫相關資訊。

	字正確率(%)
Baseline(經詞圖搜尋後辨識結果)	84.11
加上非監督式語者調適(+MLLR)後辨識結果	84.64

表三、中央社文字語料相關統計相關資訊。

新聞時間	2001 年 8 月
新聞數	14,178 則
總大小	約 709 萬字
平均每則新聞長度	約 500 字

財經、國會、影劇、醫藥衛生、體育及地方新聞，相關資訊如表三所示。

實驗的自動摘要評估的標準答案為三位國立台灣大學文學院大三以上學生分別對 200 則廣播新聞的人工轉寫內容取摘要，摘要的結果包含依據不同摘要比例選取之文句(Extraction)與重寫(Abstraction)兩種[15]。

3.2 評估方式

本論文採用二種實驗的評估方式，分別為餘弦(Cosine)評估及 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)評估[5]，我們於下面各小節作詳細說明。

3.2.1 餘弦評估

此評估方法是以計算自動摘要與人工摘要結果的相關度作為評估標準。人工摘要包含有摘錄式(Extractive)摘要(或稱為句排名方式)及非摘錄式(Non-Extractive or Abstractive)摘要(或稱為重寫摘要的方式)二種。前者可依不同的摘要比例而產生不同長度的摘要內容，如 20%、30%、50%、70% 等摘要比例；而後者則為固定長度、內容的摘要結果，無法依摘要比例作任意的更改。

假設 $A_D(m\%)$ 代表對某篇文件 D 之自動摘要 $m\%$ 摘要比例的結果、$E_{z,D}(m\%)$ 代表第 z 個人對文件 D 以摘錄方式選取 $m\%$ 摘要比例的結果、$R_{z,D}$ 代表第 z 個人對文件 D 重寫摘要結果，則對所有 $|\{D\}|$ 篇新聞的自動摘要正確率被定義為[15]：

$$ACC(m\%) = \frac{1}{|\{D\}|}\frac{1}{Z}\sum_{D\in|D|}\sum_{z=1}^{Z}\frac{SIM\left(A_D(m\%),E_{z,d}(m\%)\right)+SIM\left(A_D(m\%),R_{z,d}\right)}{2} \quad (14)$$

其中，Z 為產生人工標準答案的人數，$SIM(\cdot,\cdot)$ 為評估自動摘要與人工摘要結果的餘弦值，公式如下：

$$SIM\left(A_D(m\%),E_{z,d}(m\%)\right)=\frac{\vec{V}_{A_D(m\%)}\bullet\vec{V}_{E_{z,d}(m\%)}}{\left|\vec{V}_{A_D(m\%)}\right|\left|\vec{V}_{E_{z,d}(m\%)}\right|} \quad (15)$$

上式中 $\vec{V}_{A_D(m\%)}$ 與 $\vec{V}_{E_{z,d}(m\%)}$ 分別為自動摘要結果 $A_D(m\%)$ 與人工摘錄式摘要結果 $E_{z,D}(m\%)$ 的向量表示式。

3.2.2 ROUGE 評估

ROUGE 是召回率導向(Recall-Oriented)自動摘要評估方式[5]。此評估方法為計算自動摘要與人工摘要重疊單位元的次數，單位元可為 *N*-連語言模型(*N*-gram Language Model)、詞順序(Word Sequences)或詞對(Word Pairs)。以 *N*-連語言模型單位元的評估為例，計算方式如下所示：

$$ROUGE-N=\arg\max_{z}\frac{\sum_{S_z\in\{\text{人工標準答案}\}_z}\sum_{\text{N-連語言模型}\in S}Count_{match}(\text{N-連語言模型},S_z)}{\sum_{S_z\in\{\text{人工標準答案}\}_z}\sum_{\text{N-連語言模型}\in S}Count(\text{N-連語言模型},S_z)} \quad (16)$$

$Count_{match}(\cdot,\cdot)$ 表示 *N*-連語言模型共同出現在自動摘要與第 z 個人的人工摘要的最大次數。在本論文，當實驗以 ROUGE 為評估標準時，皆以 ROUGE-2 即雙連語言模型為比對單位，而且此評估方法僅採人工標準答案的摘錄式摘要作為摘要正確率的評估。

3.3 實驗結果

我們共採用三種基礎實驗，分別為向量空間模型(VSM)、相關評估(RM)及潛藏語意分析模型(LSA)，將它們與我們所提出之三種摘要模型：嵌入式潛藏語意分析(eLSA)、隱藏式馬可夫模型(HMM)、主題混合模型(TMM)，進行不同摘要比例下之摘要結果比較及分析。我們採用單詞為特徵單位進行自動摘要實驗，對於單詞的特徵單位，我們使用“貪婪(Greedy)演算法”(即長詞優先演算法)作為訓練語料庫及實驗語料的斷詞方法。以下將呈現在不同評估方法下各種摘要模型之摘要結果。

3.3.1 餘弦評估

在本小節中實驗之評估方法均採餘弦評估，HMM 與 TMM 數據均是使用期望值最大化演算法，自動訓練調整參數與模型所得之結果。首先，我們於表四呈現 TMM 在不同潛藏主題數下對於人工轉寫文件的自動摘要結果的影響。由表中不同潛藏主題數之結果可觀察出潛藏主題數為 64 時，低摘要比例的摘要結果之正確率為較高。接著，採用表四中得到最佳摘要結果的潛藏主題數之 TMM 模型(主題個數 64)與其他模型進行比較。

表五至表七為同樣以詞為特徵單位，在不同摘要比例下的摘要結果。以 VSM、RM、LSA 三種基礎實驗來看，VSM 的表現結果為最佳而 LSA 則明顯地較低。經由實驗觀察，在低摘要比例(例如 20%)的情況下，TMM 不論在人工轉寫或自動轉寫的文件摘要上，都呈現出優於其他摘要模型的結果，且於不同摘要比例下 TMM 及 HMM 結果均較 RM 與 LSA 高，而在高摘要比例下 eLSA 亦有不錯的結果。雖然我們所提出的 embedded LSA 模型並沒有很突出的實驗結果，但是其結果明顯較傳統的 LSA 模型高出許多，而且於高摘要比例時亦有很好的表現。進一步比較表五(人工轉寫)與表六(自動轉寫(Baseline))、表七(自動轉寫(+MLLR))之結果，發現自動轉寫的部分所有模型的正確率均下降許多，這可歸因於兩個原因：其一為語音辨識所產生的辨識錯誤對摘要結果正確率的影響；其二，為語音辨識結果(自動轉寫)的斷句與標準答案的文句不一致，由於自動轉寫的斷句方式主要是以靜音(Silence)的長度來判斷，因此可能會與標準答案的文句有所差異，而導致摘要正確率的下降。

表四、主題混合模型使用不同潛藏主題個數之比較，使用餘弦評估及人工轉寫。

摘要比例	4	8	16	32	64
20%	0.4686	0.4686	0.4688	0.4663	0.4713
30%	0.5262	0.5270	0.5270	0.5275	0.5299
50%	0.6264	0.6296	0.6271	0.6290	0.6292
70%	0.6966	0.6960	0.6961	0.6961	0.6964

表五、各種摘要模型比較之比較，使用餘弦評估及人工轉寫。

摘要比例	VSM	RM	LSA	eLSA	HMM	TMM
20%	0.4581	0.4581	0.3437	0.4385	0.4688	0.4713
30%	0.5201	0.5098	0.4409	0.5034	0.5273	0.5299
50%	0.6313	0.6185	0.5964	0.6289	0.6257	0.6292
70%	0.6964	0.6919	0.6762	0.7011	0.6962	0.6964

表六、各種摘要模型比較之比較，使用餘弦評估及自動轉寫(Baseline)。

摘要比例	VSM	RM	LSA	eLSA	HMM	TMM
20%	0.3687	0.3686	0.2913	0.3528	0.3668	0.3694
30%	0.4099	0.4106	0.3454	0.3969	0.4037	0.4055
50%	0.5251	0.5017	0.4778	0.5061	0.5162	0.5150
70%	0.5763	0.5556	0.5453	0.5710	0.5739	0.5735

表七、各種摘要模型比較之比較，使用餘弦評估及自動轉寫(+MLLR)。

摘要比例	VSM	RM	LSA	eLSA	HMM	TMM
20%	0.3734	0.3708	0.2987	0.3580	0.3727	0.3743
30%	0.4231	0.4198	0.3523	0.4012	0.4155	0.4142
50%	0.5276	0.5060	0.4838	0.5110	0.5178	0.5176
70%	0.5789	0.5615	0.5463	0.5735	0.5745	0.5743

表八、各種摘要模型比較之比較，使用 Rouge-2 評估及人工轉寫。

摘要比例	VSM	RM	LSA	eLSA	HMM	TMM
20%	0.3892	0.4062	0.2701	0.4010	0.4390	0.4438
30%	0.4544	0.4713	0.3843	0.4740	0.4874	0.4927
50%	0.6094	0.6284	0.5901	0.6601	0.6271	0.6298
70%	0.7280	0.7516	0.7035	0.7736	0.7413	0.7400

表九、各種摘要模型比較之比較，使用 Rouge-2 評估及自動轉寫(Baseline)。

摘要比例	VSM	RM	LSA	eLSA	HMM	TMM
20%	0.2463	0.2519	0.2006	0.2719	0.2632	0.2685
30%	0.2638	0.2780	0.2320	0.2781	0.2757	0.2785
50%	0.3965	0.3949	0.3608	0.4156	0.4056	0.4031
70%	0.4666	0.4647	0.4419	0.4931	0.4733	0.4710

表十、各種摘要模型比較之比較，使用 Rouge-2 評估及自動轉寫(+MLLR)。

摘要比例	VSM	RM	LSA	eLSA	HMM	TMM
20%	0.2576	0.2626	0.2035	0.2791	0.2756	0.2753
30%	0.2818	0.2963	0.2359	0.2886	0.2878	0.2859
50%	0.4033	0.4019	0.3686	0.4218	0.4114	0.4106
70%	0.4685	0.4703	0.4460	0.4958	0.4763	0.4750

3.3.2 Rouge-2 評估

在本小節中實驗之評估方法均採 Rouge-2 評估，表八至表十為此評估方法下實驗所得之結果，由表中三種基礎實驗結果來看，以 RM 方法的摘要正確率為最佳，此與 3.3.1 小節所呈現之結果略有不同，但是 VSM 亦有不錯的表現而 LSA 仍明顯地有較低的正確率。實驗結果顯示，不論在人工轉寫或是自動轉寫上，於不同摘要比例時我們所提之三種摘要模型均有顯著的表現，特別是在自動轉寫上對於不同摘要比例而言，TMM、HMM 與 eLSA 結果幾乎都較基礎實驗的結果高。此外，我們亦可發現在高摘要比例下，eLSA 的結果不僅優於傳統的 LSA 模型，其摘要正確率甚至為所有模型結果之中最高。因此，由 3.3.1 小節與本小節之實驗結果可證明 eLSA 於高摘要比例下有極佳的表現，而 TMM 及 HMM 對於摘要正確率亦有顯著之提昇。

另一方面，經由表八(人工轉寫)與表九(自動轉寫(Baseline))、表十(自動轉寫(+MLLR))之結果比較，可以觀察出所提出的三種摘要模型於自動轉寫上有優於在人工轉寫上之表現，由此可證明所提出的三種模型較其他模型更適於用自動轉寫的應用，即適用於語音文件上。

3.4 HMM 與 TMM 的進一步探討

對於 HMM 與 TMM 來說，每個被摘要的文件 D 中皆含有文句模型 S_i 的資訊，此模型 S_i 的字詞資訊可

表十一、HMM 文句移除實驗
（評估方式：餘弦評估，特徵單位：詞）

摘要比例	人工轉寫		自動轉寫(Baseline)		自動轉寫(+MLLR)	
	HMM	HMM SR	HMM	HMM SR	HMM	HMM SR
20%	0.4688	0.4712	0.3668	0.3664	0.3727	0.3725
30%	0.5273	0.5282	0.4037	0.4041	0.4155	0.4098
50%	0.6257	0.6301	0.5162	0.5139	0.5178	0.5170
70%	0.6962	0.6979	0.5739	0.5729	0.5745	0.5771

表十二、HMM 文句移除實驗
（評估方式：Rouge-2 評估，特徵單位：詞）

摘要比例	人工轉寫		自動轉寫(Baseline)		自動轉寫(+MLLR)	
	HMM	HMM SR	HMM	HMM SR	HMM	HMM SR
20%	0.4390	0.4381	0.2632	0.2669	0.2756	0.2718
30%	0.4874	0.5011	0.2757	0.2747	0.2878	0.2841
50%	0.6271	0.6434	0.4056	0.4110	0.4114	0.4146
70%	0.7413	0.7595	0.4733	0.4811	0.4763	0.4856

表十三、TMM 文句移除實驗
（評估方式：餘弦評估，特徵單位：詞）

摘要比例	人工轉寫		自動轉寫(Baseline)		自動轉寫(+MLLR)	
	TMM	TMM SR	TMM	TMM SR	TMM	TMM SR
20%	0.4713	0.4719	0.3694	0.3665	0.3743	0.3725
30%	0.5299	0.5319	0.4055	0.4039	0.4142	0.4100
50%	0.6292	0.6306	0.5150	0.5142	0.5176	0.5189
70%	0.6964	0.6971	0.5735	0.5732	0.5743	0.5765

表十四、TMM 文句移除實驗
（評估方式：Rouge-2 評估，特徵單位：詞）

摘要比例	人工轉寫		自動轉寫(Baseline)		自動轉寫(+MLLR)	
	TMM	TMM SR	TMM	TMM SR	TMM	TMM SR
20%	0.4438	0.4390	0.2685	0.2689	0.2753	0.2718
30%	0.4927	0.5037	0.2785	0.2749	0.2859	0.2841
50%	0.6298	0.6431	0.4031	0.4108	0.4106	0.4147
70%	0.7400	0.7579	0.4710	0.4817	0.4750	0.4851

以由文件 D 中去除，此稱做文句移除（Sentence Removal, SR），再進行機率值 $P(D|S_i)$ 的估測，如 2.2 小節中所提。所以，我們進一步針對 HMM 及 TMM 進行文句移除實驗，藉以觀察文句移除對於摘要結果是否有所影響或是提昇。表十一、表十二分別為 HMM 未採用文句移除之結果與進行文句移除後結果於不同評估方式下之對照，在人工轉寫文件上，由二種評估方式的結果均可看出文句移除有助於自動摘要正確率的提昇，然而在自動轉寫文件上，僅有 Rouge-2 評估的結果可明顯觀察出文句移除對摘要結果的提昇，這可能是因辨識錯誤及斷句方式的不同，抵銷了其結果使得正確率無明顯的改進，

但是可以發現在辨識率上升時(+MLLR 較 Baseline 好)，其結果有所提高。表十三、表十四為 TMM 未採用文句移除之結果與進行文句移除後結果於不同評估方式下之對照，經由結果亦可觀察出於人工轉寫文件上文句移除有助於自動摘要正確率的提昇，雖然在自動轉寫文件上可能因辨識所產生的問題而無法看出其效果，但是實驗結果仍然可以證明文句移除確實有助於文件自動摘要正確率的提昇。

4. 結論與未來展望

本論文對於自動文件摘要提出三種摘要模型：在逐字比對方式，提出以隱藏式馬可夫模型(Hidden Markov Model, HMM)為摘要模型；在概念比對方式，提出以嵌入式潛藏語意分析(embedded LSA, eLSA)、主題混合模型(Topical Mixture Model, TMM)作為摘要模型。我們在中文語音廣播新聞語料庫上實作了一系列的實驗，並分別採用二種評估方式進行結果的分析。由使用人工轉寫及使用自動轉寫來從事摘要實驗的結果證明我們所提出的摘要模型具有顯著的效果，說明我們所提出的摘要模型非常適用於語音文件上。同時，我們也進行 HMM 與 TMM 文句移除的實驗，證明文句移除確實有助於文件自動摘要正確率的提昇。

在本論文，我們採用以詞為特徵單位進行摘要結果的分析與比較，但根據我們近年來在中文語音文件檢索的研究發現[3]，使用音節(Syllable)或字(Character)為索引特徵往往可以較以詞為索引特徵有更好的檢索表現。若以所帶資訊的豐富程度來看，詞所含的資訊量是最多的，其次是字，然後為音節[12]。但是以語音辨識正確率來講，音節的辨識正確率較其他二種特徵單位高，其次是字，然後為詞。因此，我們相信除了使用單詞為特徵單位進行自動摘要外，亦可使用以單字、雙字、單音節、雙音節為特徵單位進行實驗，探討不同特徵單位對自動摘要的正確率影響，我們相信可對於語音文件的摘要正確率有明顯的提高。另外，針對隱藏式馬可夫模型，其將文件中每一字句視為一個生成模型，由於一字句所含有的資訊量較不足，因此可以嘗試將每一字句模型作擴充(Expansion)以提升摘要正確率。

在摘要模型上，未來的研究方向我們將考慮以更多可使用的資訊來提高摘要的正確率，像是動態結合屬性資訊(如新聞開頭、結論段落為重要字句)、結合更多自然語言方面的資訊，如詞性(Part-of-Speech, POS)、結合語音聲學上特性，如音高、能量等，或是探討混合不同摘要模型的方法。此外，自動摘要的結果亦可應用於文件分類器及資訊檢索上，未來我們也將朝著這個方向努力。以期進一步提升自動摘要結果，也對於其他相關領域有所助益。

5. 參考文獻

[1] B. Chen, “Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval,” accepted for publication in *Pattern Recognition Letters*, 2005.

[2] B. Chen, H.M. Wang, L.S. Lee, “A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents,” *ACM Transactions on Asian Language Information Processing*, vol. 3, no. 2, June 2004, pp. 128-145.

[3] B. Chen, H.M. Wang and L.S. Lee, ”Discriminating Capability of Syllable-based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese,” *IEEE Trans. on Speech and Audio Processing*, vol.10, no5, 2002, pp. 303-314.

[4] B. Chen, Y.T. Chen, C.H. Chang, H.B. Chen, “Speech Retrieval of Mandarin Broadcast News via Mobile Devices,” the 9th *European Conference on Speech Communication and Technology*, September 2005.

[5] C.Y. Lin, “ROUGE: Recall-oriented Understudy for Gisting Evaluation,” 2003, http://www.isi.edu/~cyl/ROUGE/.

[6] Dempster, A.P., Laird, N. M., Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society B, Vol. 39, No. 1, 1-38.

[7] G. Salton and M. E. Lesk, “Computer evaluation of indexing and text processing,” *Journal of the ACM*, vol. 15, no. 1, pp. 8-36, 1968.

[8] G.W. Furnas, S. Deerwester, S.T. Dumais, T.K Landauer., R. Harshman, L.A. Streeter and K.E. Lochbaum, “Information retrieval using a singular value decomposition model of latent semantic structure,” in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 1988, pp. 465-480.

[9] I. Mani and M. T. Maybury, Eds., *Advances in Automatic Text Summarization*. Cambridge. MA: MIT Press, 1999.

[10] J. Goldstein, M. Kantrowitz, V. Mittal and J Carbonell, “Summarizing text documents: sentence selection and evaluation metrics,” in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 1999, pp. 121-128.

[11] Lin-shan Lee and Berlin Chen, “Spoken Document Understanding and Organization,” *IEEE Signal Processing Magazine (IEEE SPM)*, Vol. 22, No. 5, Sept. 2005, pp. 42-60.

[12] L.S. Lee, Y. Ho, J.F. Chen, S.C. Chen, ”Why is the Special Structure of the Language Important for Chinese Spoken Language Processing -Examples on Spoken Document Retrieval, Segmentation, and Summarization,” the 8th *European Conference on Speech Communication and Technology*, September 2003.

[13] S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori,

"Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 4, pp. 401-408, 2004.

[14] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 2001, pp. 19-25.

[15] Y. Ho, *An initial study on automatic summarization of Chinese spoken documents*. Master Thesis, National Taiwan University, July 2003.

[16] Central News Agency, http://210.69.89.224/search/hypage.cgi?HYPAGE=login.htm